IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

BOARD OF PATENT APPEALS AND INTERFERENCES

In re patent application of:
William Scott Spangler

Serial No.: 09/669,680 Group Art Unit: 2176

Filed: September 26, 2000 Examiner: Kyle Stork

For: A METHOD FOR ADAPTING A K-MEANS TEXT CLUSTERING TO EMERGING DATA

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

APPELLANTS' APPEAL BRIEF

Sirs:

Appellant respectfully appeals the final rejection of claims 8, 11-15, 17-23, and 26-29, in the Office Action dated March 26, 2007. A Notice of Appeal was timely filed on June 26, 2007.

I. REAL PARTY IN INTEREST

The real party in interest is International Business Machines Corporation, Armonk, New York, assignee of 100% interest of the above-referenced patent application.

II. RELATED APPEALS AND INTERFERENCES

There are no other appeals or interferences known to Appellants, Appellants' legal representative or Assignee which would directly affect or be directly affected by or have a bearing on the Board's decision in this appeal.

III. STATUS OF CLAIMS

Claims 8, 11-15, 17-23, and 26-29, all the claims pending in the application are under appeal, and stand rejected under 35 U.S.C. §103(a) as being unpatentable over Lantrip, et al., (U.S. Patent No. 6,298,174) hereinafter referred to as "Lantrip," in view of Ruocco, et al. (U.S. Patent No. 5,864,855) hereinafter referred to as "Ruocco."

IV. STATUS OF AMENDMENTS

The present application was filed on September 26, 2000. A non-final office action was issued on March 25, 2004. An amendment was filed under 37 C.F.R. §1.111 on June 23, 2004. A final office action was issued November 17, 2005. An after-final amendment was filed under 37 C.F.R. §1.116 on January 13, 2005. A request for continued examination was filed on February 14, 2005. A Preliminary Amendment was filed on April 19, 2005. A non-final office action was mailed on April 18, 2005 and received on April 21, 2005. On May 17, 2005 an amendment was filed under 37 C.F.R. §1.111. A notice of non-compliant amendment was mailed July 26, 2005. In response,

on August 24, 2005, another amendment was filed under 37 C.F.R. §1.111. A notice of abandonment was mailed October 13, 2006. A petition to withdraw the holding of abandonment was filed on November 1, 2006, due to USPTO error. The petition was granted January 22, 2007.

A final rejection, in response to the communication filed on November 1, 2006, was issued March 26, 2007. An after-final Amendment was filed on May 25, 2007 to correct a spelling error in claim 8. An Advisory Action dated June 25, 2007 indicated that, the Amendment filed on May 25, 2007 would not be entered, and that the rejections of claims would remain. Thus, the claims shown in the appendix are shown in their amended form as of the August 24, 2005 Amendment. Please note that once the claims are allowed, the Applicants will file a 37 C.F.R. §1.312 amendment to correct the known spelling error in claim 8 (i.e., in order to change "mode" to "model" on line 13 of claim 8) or any other such typographical errors (e.g., claim 26 will be amended to depend from claim 23 rather than cancelled claim 25; page 8, line 17 will be amended to change "mode" to "model", etc.).

V. SUMMARY OF CLAIMED SUBJECT MATER

A. Independent Claim 8

Claim 8 reflects an embodiment of a system of the invention for clustering documents in datasets. Specifically, claim 8 defines "a system for clustering documents in datasets." Such a system is described at various points in the specification, for example, page 5 lines 15-16, indicates that the invention provides "a structure and method of clustering documents in datasets". Figure 2 illustrates a schematic architectural diagram of such an embodiment. The specification further defines each of the several individual features of this system.

One feature of the system is storage with first and second datasets. Claim 8 defines this feature as "a storage having a first dataset and a second dataset". This feature is described at various points in the specification. For example, page 7, lines 1-3,

describe this feature as follows: "a storage having a first dataset and a second data set". This feature is further described, for example, at page 8, lines 3-6 and illustrated as items 200 and 202 in Figure 2.

Another feature of the system is "a cluster generator operative to cluster first documents in the first dataset to produce first document classes." Claim 8 defines this feature as follows: "a cluster generator operative to cluster first documents in said first dataset and produce first document classes." This feature is described at various points in the specification. For example, on page 9, lines 6-7, this feature is described as follows: "a K-means cluster generator 222 clusters (i.e. partitions) the documents in the first dataset T1 based on the T1-D1 vector space model. The clustering algorithm "K-means" is one of the most popular procedures for automatic classification of data when no classification is known." This is shown in Figure 1, item 104 and Figure 2, items 222-224.

Another feature of the invention is a centroid seed generator operative to generate centroid seeds based on the first document classes. Claim 8 defines this feature as follows: "a centroid seed generator operative to generate centroid seeds based on said first document classes." This feature is described at various points in the specification. For example, on page 5, lines 17-18, this feature is described as "creating centroid seeds based on the first document classes." Additionally, on page 10, line 20-page ll, line 5, this feature is described as follows: "A centroid seed generator 220 then creates centroid seeds based on the mean of each of the classes from the T2-D1 vector space model 214 as classified by the classifier 218 (based on the K-classes from the T1-D1 cluster) and inputs the centroid seeds to the K-means cluster generator 222, as shown in item 114." This feature is shown as item 220 of Figure 2.

Another feature of the invention is a dictionary generator adapted to generate a first dictionary of most common words in the first dataset. Claim 8 defines this feature as follows: "a dictionary generator adapted to generate a first dictionary of most common words in said first dataset." This feature is described at various points in the specification. For example, page 8, lines 9-11 describes this feature as follows: "the

invention begins by generating a first dictionary 206, D1, of frequently used words from dataset T1 200 using a dictionary generator 204". This is shown as item 100 of Figure 1 and items 2-4-206 of Figure 2.

Another feature of the invention is a vector space model generator adapted to generate a first vector space model by counting, for each word in the first dictionary, a number of the first documents in which the word occurs. Claim 8 defines this feature as follows: "a vector space model generator adapted to generate a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said word occurs." This feature is described at various points in the specification. For example, page 8, lines 15-17, describes this feature as follows: "the vector space model generator 210 counts, for each word in the first dictionary D1 206, the number of documents in which the word in question appears, to produce a T1-D1 vector space model 212". This is shown as item 102 of Figure 1 and items 210-212 of Figure 2.

Another feature of the invention is that the cluster generator, described above, clusters the documents in the first dataset based on the first vector space model. Claim 8 defines this feature as follows: "wherein said cluster generator clusters said documents in said first dataset based on said first vector space model." This feature is described at various points in the specification. For example, page 6, lines 1-7, described this feature as follows: "… clusters the first documents in the first dataset based on the first vector space model…". Additionally, page 9, lines 6-7, describes this feature as follows: "In item 104, a K-means cluster generator 222 clusters (i.e. partitions) the documents in the first set T1 based on the T1-D1 vector space model". This is as items 102-104 in Figure 1 and as items 212-222-224 in Figure 2.

Another feature of the invention is that the cluster generator, described above, also clusters the second documents in the second dataset using the centroid seeds, such that the second dataset has a similar, based on the centroid seeds, clustering to that of the first dataset. Claim 8 defines this feature as follows: "wherein said cluster generator clusters second documents in said second dataset using said centroid seeds, such that said second dataset has a similar, based on said centroid seeds, clustering to that of said first

dataset." This feature is described at various points in the specification. For example, page 10, line 1-11, provides: "Next, in item 106, for the second test dataset T2 202, the vector space model generator 210 generates a T2-D1 vector space model 214, by counting for each word, the number of documents in the second dataset T2 202 in which the word in the D1 dictionary 206 appears. Following that, in item 108, the classifier 218 classifies the document in the t2-D1 vector space model 214 by finding for each document in T2 202, the nearest centroid (based on the K-classes of the T1-D1 cluster 224) to that document using the distance metric (e.g., Cosine) from item 104. In other words, the invention classifies the documents within the T2 data 202, using the classes produced during the generation of the T1-D1 cluster 224 to make the clustering of the T2 data 202 similar to the clustering on the T1 data 200". Additionally, page 10, line 20 – page 11, line 2 describes this feature as follows: "A centroid seed generator 220 then creates centroid seeds based on the mean of each of the classes from the T2-D1 vector space model 214 as classified by the classifier 218 (based on the K-classes from the T1-D1 cluster) and inputs the centroid seeds to the K-means cluster generator 222, as shown in item 114. The initial centroid (seed) for each class is found by summing up the columns of all examples in the class and dividing these values by the number of elements in the class. Centroid seeds are used to generate the initial clustering which is then optimized using the K-means approach." This is shown as items 118 of Figure 1 and 210-228 of Figure 2.

Another feature of the invention is second dataset comprises a new, but related, based on the centroid seeds, dataset different than the first dataset. Claim 8 defines this feature as follows: "second dataset comprises a new, but related, based on said centroid seeds, dataset different than said first dataset." This feature is described at various points in the specification. For example, page 8, lines 3-6 describes this feature as follows: "it is assumed that an initial text dataset, T1 (e.g., January helpdesk data), is classified first, followed by a new, but related text dataset, T2 (e.g., February helpdesk data) which should be classified similarly, but should also be indexed to highlight emerging trends". Page 11, lines 16-19, further provides that the "invention intentionally biases the

algorithm towards the previous classification centroids. Thus, the invention directs the K-means solution towards the original classification (January) without preventing it from adjusting that classification in February as the data determines."

B. Independent Claim 15

Claim 15 reflects an embodiment of a method of the invention for clustering documents in datasets. Specifically, claim 15 defines "A method of clustering documents in a first dataset having first documents and a related second dataset having second documents." Such a method is described at various points in the specification, for example, page 5 lines 15-16, indicates that the invention provides "a structure and method of clustering documents in datasets". Figure 1 is a flowchart illustrating the method. The specification further defines each of the several individual features of this method.

One feature of the method is clustering the first documents to produce first document classes. Claim 15 defines this feature as follows: "clustering said first documents to produce first document classes." This feature is described at various points in the specification. For example, on page 9, lines 6-7, this feature is described as follows: "a K-means cluster generator 222 clusters (i.e. partitions) the documents in the first dataset T1 based on the T1-D1 vector space model. The clustering algorithm "K-means" is one of the most popular procedures for automatic classification of data when no classification is known." This is shown as item 104 of Figure 1 and items 222-224 of Figure 2.

Another feature of the invention is generating a vector space model of the second documents. Claim 15 defines this feature as follows: "generating a vector space model of said second documents." This feature is described at various points in the specification, for example page 10, lines 1-4 describes this feature as follows: "in item 106, for the second text dataset T2 202, the vector space model generator 210 generates a T2-D1 vector space model 214, by counting, for each word, the number of documents in the

second dataset T2 202 in which the word in the D1 dictionary 206 appears". This is shown as item 106 of Figure 1 and items 214 of Figure 2.

Another feature of the invention is classifying the vector space model of the second documents using the first document classes to produce a classified vector space model. Claim 15 defines this feature as follows: "classifying said vector space model of said second documents using said first document classes to produce a classified vector space model." This feature is described at various points in the specification. For example, page 10, lines 8-11, describes this feature as follows: "in item 108, the classifier 218 classifies the documents in the T2-D1 vector space model 214 by finding for each document T2 202, the nearest centroid (based on the K-classes of the T1-D1 cluster 224) to that document using the distance metric (e.g., Cosine) from item 104. In other words, the invention classifies the documents within the T2 data 202, using the classes produced during the generation of the T1-D1 cluster 224 to make the clustering of the T2 data 202 similar to the clustering on the T1 data 200". This is shown as item 108 in Figure 1 and as items 214-218 of Figure 2.

Another feature of the invention is determining a mean of vectors in each class in the classified vector space model to produce centroid seeds. Claim 15 defines this feature as follows: "determining a mean of vectors in each class in said classified vector space model to produce centroid seeds." This feature is described at various points in the specification. For example, page 10, line 20 – page 11, line 2, describes this feature as follows: "A centroid seed generator 220 then creates centroid seeds based on the mean of each of the classes from the T2-D1 vector space model 214 as classified by the classifier 218 (based on the K-classes from the T1-D1 cluster) and inputs the centroid seeds to the K-means cluster generator 222, as shown in item 114". This is shown as item 114 in Figure 1 and item 220 of Figure 2.

Another feature of the invention is clustering the second documents using the centroid seeds, such that the second dataset has a similar, based on the centroid seeds, clustering to that of the first dataset. Claim15 defines this feature as follows: "clustering said second documents using said centroid seeds, such that said second dataset has a

similar, based on said centroid seeds, clustering to that of said first dataset." This feature is described at various points throughout the specification. For example, page 10, line 1-11, provides: "Next, in item 106, for the second test dataset T2 202, the vector space model generator 210 generates a T2-D1 vector space model 214, by counting for each word, the number of documents in the second dataset T2 202 in which the word in the D1 dictionary 206 appears. Following that, in item 108, the classifier 218 classifies the document in the t2-D1 vector space model 214 by finding for each document in T2 202, the nearest centroid (based on the K-classes of the T1-D1 cluster 224) to that document using the distance metric (e.g., Cosine) from item 104. In other words, the invention classifies the documents within the T2 data 202, using the classes produced during the generation of the T1-D1 cluster 224 to make the clustering of the T2 data 202 similar to the clustering on the T1 data 200". Additionally, page 10, line 20 – page 11, line 2 describes this feature as follows: "A centroid seed generator 220 then creates centroid seeds based on the mean of each of the classes from the T2-D1 vector space model 214 as classified by the classifier 218 (based on the K-classes from the T1-D1 cluster) and inputs the centroid seeds to the K-means cluster generator 222, as shown in item 114. The initial centroid (seed) for each class is found by summing up the columns of all examples in the class and dividing these values by the number of elements in the class. Centroid seeds are used to generate the initial clustering which is then optimized using the K-means approach." This is shown as items 118 of Figure 1 and 210-228 of Figure 2.

Another feature of the invention is second dataset comprises a new, but related, based on the centroid seeds, dataset different than the first dataset. Claim 15 defines this feature as follows: "second dataset comprises a new, but related, based on said centroid seeds, dataset different than said first dataset." This feature is described at various points in the specification. For example, page 8, lines 3-6 describes this feature as follows: "it is assumed that an initial text dataset, T1 (e.g., January helpdesk data), is classified first, followed by a new, but related text dataset, T2 (e.g., February helpdesk data) which should be classified similarly, but should also be indexed to highlight emerging trends". Page 11, lines 16-19, further provides that the "invention intentionally biases the

algorithm towards the previous classification centroids. Thus, the invention directs the K-means solution towards the original classification (January) without preventing it from adjusting that classification in February as the data determines." This is shown in Figures 1 and 2.

Another feature of the invention is that there are at least two vector space models generated: a first vector space model and a second vector space model (see page 8, line 17 and page 10, line 2). Additionally, process of clustering the first document in the first dataset is accomplished by forming a first dictionary of most common words in the first dataset and generating the first vector space model by counting, for each word in the first dictionary, a number of the first document in which the word occurs. More specifically, claim 15 defines this feature as follows: "wherein said vector space model comprises a second vector space model and said clustering of said first documents in said first data comprises: forming a first dictionary of most common words in said first dataset and generating a first vector space model by counting, for each word in the first dictionary, a number of said first document in which said word occurs." This feature is described at various points in the specification. For example, page 8, lines 9-17, describes this feature as follows: "the invention begins by generating a first dictionary 206, D1, of frequently used words form dataset T1 200 using a dictionary generator 204. The most frequently occurring words in the corpus make up the dictionary. This reduced set of words will be used to compose a simple description of each document in the corpus. The number of words to be included in the dictionary is user specified parameter. Then, the vector space model generator 210 counts, for each word in the first dictionary D1 206, the number of documents in which the word in question appears, to produce a T1-D1 vector space mode 212." This is shown in Figures 1 and 2.

Another feature of the invention is that the clustering of the first documents in the first dataset, described above, is based on the first vector space mode. Claim 15 defines this feature as follows: "clustering of said first documents in said first dataset based on said first vector space mode." This feature is described at various points in the specification. For example, page 9, lines 6-9, describes this feature as follows: "In item

104, a K-means cluster generator 222 clusters (i.e., partitions) the documents in the first dataset T1 based on the T1-D1 vector space model". This is as item 104 of Figure 1 and items 212-222-224 of Figure 2.

C. Independent Claim 20

Claim 20 reflects another embodiment of a method of the invention for clustering documents in related datasets. Specifically, claim 20 defines "A method of clustering documents in related datasets." Such a method is described at various points in the specification. For example, page 5, lines 15-16, indicates that the invention provides "a structure and method of clustering documents in datasets". Page 8, lines 3-6, further provide it is assumed that an initial dataset T1 (e.g., January helpdesk data) is classified first, followed by a new, but related text dataset, T2 (e.g., February helpdesk data) which should be classified similarly, but should also be indexed to highlight emerging trends. Figure 1 is a flowchart illustrating this method. The specification further defines each of the several individual features of this method.

One feature of the invention is the formation of a first dictionary of most common words in a first dataset. Claim 20 defines this feature as follows: "forming a first dictionary of most common words in said first dataset." This feature is described at various points in the specification. For example, page 8, lines 9-14, describes this feature as follows: "the invention begins by generating a first dictionary 206, D1, of frequently used words form dataset T1 200 using a dictionary generator 204. The most frequently occurring words in the corpus make up the dictionary. This reduced set of words will be used to compose a simple description of each document in the corpus. The number of words to be included in the dictionary is user specified parameter." This is shown in as item 100 of Figure 1 and as 200-204-206 of Figure 2.

Another feature of the invention is generating a first vector space model by counting, for each word in the first dictionary, a number of the first documents in which the word occurs. Claim 20 defines this feature as follows: "generating a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said word occurs." This feature is described at various points in the

specification. For example, page 8, lines 15-17, describes this feature as follows: "the vector space model generator 210 counts, for each word in the first dictionary D1 206, the number of documents in which the word in question appears, to produce a T1-D1 vector space mode 212". This is shown as item 102 of Figure 1 and items 210-212 of Figure 2.

Another feature of the invention is clustering the first documents in said first dataset based on said first vector space model to produce first document classes. Claim 20 defines this feature as follows: "clustering said first documents in said first dataset based on said first vector space model to produce first document classes." This feature is described at various points in the specification. For example, page 9, lines 6-7, describes this feature as follows: "In item 104, a K-means cluster generator 222 clusters (i.e. partitions) the documents in the first dataset T1 based on the T1-D1 vector space model. The clustering algorithm "K-means" is one of the most popular procedures for automatic classification of data when no classification is known". This is shown as item 104 in Figure 1 and as items 222-224 of Figure 2.

Another feature of the invention is generating a second vector space model by counting, for each word in the first dictionary, a number of the first documents in which the word occurs. Claim 20 defines this feature as follows: "generating a second vector space model by counting, for each word in said first dictionary, a number of said second documents in which said word occurs." This feature is described at various points in the specification. For example, page 10, lines 1-4, describes this feature as follows: "for the second text dataset T2 202, the vector space model generator 210 generates a T2-D1 vector space model 214, by counting, for each word, the number of documents in the second dataset T2 202 in which the word in the D1 dictionary 206 appears". This is shown as item 106 in Figure 1 and as items 210-214 of Figure 2.

Another feature of the invention is classifying the second document in the second vector space model using the first document classes to produce a classified vector space model. Claim 20 defines this feature as follows: "classifying said second documents in said second vector space model using said first document classes to produce a classified

second vector space model." This feature is described at various points in the specification. For example, page 10, lines 5-11, describes this feature as follows: "in item 108, the classifier 218 classifies the documents in the T2-D1 vector space model 214 by finding for each document in T2 202, the nearest centroid (based on the K-classes of the T1-D1 cluster 224) to that document using the distance metric (e.g., Cosine) from item 104. In other words, the invention classifies the documents within the T2 data 202, using the classes produced during the generation of the T1-D1 cluster 224 to make the clustering of the T2 data 202 similar to the clustering on the T1 data 200". This is shown as item 108 in Figure 1 and item 218 in Figure 2.

Another feature of the invention is determining a mean of vectors in each class in the classified vector space model to produce centroid seeds. Claim 20 defines this feature as follows: "determining a mean of vectors in each class in said classified vector space model to produce centroid seeds." This feature is described at various points in the specification. For example, page 10, line 20 – page 11, line 2 describes this feature as follows: "A centroid seed generator 220 then creates centroid seeds based on the mean of each of the classes from the T2-D1 vector space model 214 as classified by the classifier 218 (based on the K-classes from the T1-D1 cluster) and inputs the centroid seeds to the K-means cluster generator 222, as shown in item 114". This is shown as item 114 of Figure 1 and item 222 in Figure 2.

Another feature of the invention is clustering second documents in a second dataset using the centroid seeds, such that the second dataset has a similar, based on the centroid seeds, clustering to that of the first dataset. Claim 20 defines this feature as follows: "clustering second documents in a second dataset using said centroid seeds, such that said second dataset has a similar, based on said centroid seeds, clustering to that of said first dataset". This feature is described at various points in the specification. For example, page 10, line 1-11, provides: "Next, in item 106, for the second test dataset T2 202, the vector space model generator 210 generates a T2-D1 vector space model 214, by counting for each word, the number of documents in the second dataset T2 202 in which the word in the D1 dictionary 206 appears. Following that, in item 108, the classifier 218

classifies the document in the t2-D1 vector space model 214 by finding for each document in T2 202, the nearest centroid (based on the K-classes of the T1-D1 cluster 224) to that document using the distance metric (e.g., Cosine) from item 104. In other words, the invention classifies the documents within the T2 data 202, using the classes produced during the generation of the T1-D1 cluster 224 to make the clustering of the T2 data 202 similar to the clustering on the T1 data 200". Additionally, page 10, line 20 – page 11, line 2 describes this feature as follows: "A centroid seed generator 220 then creates centroid seeds based on the mean of each of the classes from the T2-D1 vector space model 214 as classified by the classifier 218 (based on the K-classes from the T1-D1 cluster) and inputs the centroid seeds to the K-means cluster generator 222, as shown in item 114. The initial centroid (seed) for each class is found by summing up the columns of all examples in the class and dividing these values by the number of elements in the class. Centroid seeds are used to generate the initial clustering which is then optimized using the K-means approach." This is shown as items 118 of Figure 1 and 210-228 of Figure 2.

Another feature of the invention is that the second dataset comprises a new, but related, based on the centroid seeds, dataset different than the first dataset. Claim 20 defines this feature as follows: "second dataset comprises a new, but related, based on said centroid seeds, dataset different than said first dataset." This feature is described at various points in the specification. For example, page 8, lines 3-6 describes this feature as follows: "it is assumed that an initial text dataset, T1 (e.g., January helpdesk data), is classified first, followed by a new, but related text dataset, T2 (e.g., February helpdesk data) which should be classified similarly, but should also be indexed to highlight emerging trends". Page 11, lines 16-19, further provides that the "invention intentionally biases the algorithm towards the previous classification centroids. Thus, the invention directs the K-means solution towards the original classification (January) without preventing it from adjusting that classification in February as the data determines." This is shown in Figures 1 and 2.

D. Independent Claim 23

Claim 23 reflects an embodiment of a program device readable by machine embodying a program of instructions executable by the machine to perform a method of clustering documents in datasets. That is, as discussed on page 13, lines 3-5, the the invention can be implemented as a computer program running on top of a Java virtual machine. Further details, of how the invention would be implemented on a computer proram stored on a storage device of a computer system are provided on page 16, lines 6-20. Page 5, lines 15-16, indicates that the invention provides "a structure and method of clustering documents in datasets". Figure 1 is a flowchart illustrating this method. The specification further defines each of the several individual features of this method.

One feature of this method is clustering first documents in a first dataset to produce first document classes. Claim 23 defines this feature as follows: "clustering first documents in a first dataset to produce first document classes." This feature is described at various points in the specification. For example, on page 9, lines 6-7, this feature is described as follows: "a K-means cluster generator 222 clusters (i.e. partitions) the documents in the first dataset T1 based on the T1-D1 vector space model. The clustering algorithm "K-means" is one of the most popular procedures for automatic classification of data when no classification is known." This is shown as item 104 of Figure 1 and items 222-224 of Figure 2.

Another feature of the invention is creating centroid seeds based on the first document classes. Claim23 defines this feature as follows: "creating centroid seeds based on said first document classes." This feature is described at various points in the specification. For example, page 10, line 20 – page 11, line 2, describes this feature as follows: "A centroid seed generator 220 then creates centroid seeds based on the mean of each of the classes from the T2-D1 vector space model 214 as classified by the classifier 218 (based on the K-classes from the T1-D1 cluster) and inputs the centroid seeds to the K-means cluster generator 222, as shown in item 114". This is shown as item 114 in Figure 1 and item 220 of Figure 2.

Another feature of the invention is clustering second documents in a second dataset using the centroid seeds, such that the second dataset has a similar, based on the centroid seeds, clustering to that of the first dataset. Claim 23 defines this feature as follows: "clustering second documents in a second dataset using said centroid seeds, such that said second dataset has a similar, based on said centroid seeds, clustering to that of said first dataset." This feature is described at various points in the specification. For example, page 10, line 1-11, provides: "Next, in item 106, for the second test dataset T2 202, the vector space model generator 210 generates a T2-D1 vector space model 214, by counting for each word, the number of documents in the second dataset T2 202 in which the word in the D1 dictionary 206 appears. Following that, in item 108, the classifier 218 classifies the document in the t2-D1 vector space model 214 by finding for each document in T2 202, the nearest centroid (based on the K-classes of the T1-D1 cluster 224) to that document using the distance metric (e.g., Cosine) from item 104. In other words, the invention classifies the documents within the T2 data 202, using the classes produced during the generation of the T1-D1 cluster 224 to make the clustering of the T2 data 202 similar to the clustering on the T1 data 200". Additionally, page 10, line 20 – page 11, line 2 describes this feature as follows: "A centroid seed generator 220 then creates centroid seeds based on the mean of each of the classes from the T2-D1 vector space model 214 as classified by the classifier 218 (based on the K-classes from the T1-D1 cluster) and inputs the centroid seeds to the K-means cluster generator 222, as shown in item 114. The initial centroid (seed) for each class is found by summing up the columns of all examples in the class and dividing these values by the number of elements in the class. Centroid seeds are used to generate the initial clustering which is then optimized using the K-means approach." This is shown as items 118 of Figure 1 and 210-228 of Figure 2.

Another feature of the invention is second dataset comprises a new, but related, based on the centroid seeds, dataset different than the first dataset. Claim 23 defines this feature as follows: "second dataset comprises a new, but related, based on said centroid seeds, dataset different than said first dataset." This feature is described at various points

in the specification. For example, page 8, lines 3-6 describes this feature as follows: "it is assumed that an initial text dataset, T1 (e.g., January helpdesk data), is classified first, followed by a new, but related text dataset, T2 (e.g., February helpdesk data) which should be classified similarly, but should also be indexed to highlight emerging trends". Page 11, lines 16-19, further provides that the "invention intentionally biases the algorithm towards the previous classification centroids. Thus, the invention directs the K-means solution towards the original classification (January) without preventing it from adjusting that classification in February as the data determines."

The feature, described above, of clustering the first document in the first dataset is accomplished by forming a first dictionary of most common words in the first dataset; generating a first vector space model by counting, for each word in the first dictionary, a number of the first documents in which the word occurs; and clustering of the first documents in the first dataset based on the first vector space model. Claim 23 defines this feature as follows: "forming a first dictionary of most common words in said first dataset; generating a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said word occurs; and clustering of said first documents in said first dataset based on said first vector space model." These features are described at various points in the specification.

For example, page 8, lines 9-14, describes forming a first dictionary features as follows: "the invention begins by generating a first dictionary 206, D1, of frequently used words form dataset T1 200 using a dictionary generator 204. The most frequently occurring words in the corpus make up the dictionary. This reduced set of words will be used to compose a simple description of each document in the corpus. The number of words to be included in the dictionary is user specified parameter." This is shown in as item 100 of Figure 1 and as 200-2004-2006 of Figure 2. Page 8, lines 15-17, describes the generating a first vector space model feature as follows: "the vector space model generator 210 counts, for each word in the first dictionary D1 206, the number of documents in which the word in question appears, to produce a T1-D1 vector space mode 212". This is shown as item 102 of Figure 1 and items 210-212 of Figure 2. Finally,

page 9, lines 6-7, describes the clustering of the first documents feature as follows: "In item 104, a K-means cluster generator 222 clusters (i.e. partitions) the documents in the first dataset T1 based on the T1-D1 vector space model. The clustering algorithm "K-means" is one of the most popular procedures for automatic classification of data when no classification is known". This is shown as item 104 in Figure 1 and as items 222-224 of Figure 2.

VI. GROUNDS OF REJECTION TO BE REVIEWED ON APPEAL

The issues presented for review 8, 11-15, 17-23, and 26-29 are unpatentable under 35 U.S.C. §103(a) by Lantrip, in view of Ruocco.

VII. ARGUMENT

A. The Position in the Office Action

The Office Action states:

Claims 8, 11-15, 17-23, and 26-29 remain rejected under 35 U.S.C. 103(a) as being unpatentable over Lantrip et al. (USPN 6,298,174 B1—filing date 10/15/1999), hereinafter Lantrip, further in view of Ruocco et al.. (USPN 5,864,855—filing date 2/26/1996), hereinafter Ruocco.

Regarding independent claim 8, Lantnip discloses a method of clustering documents in datasets (in col. 2, lines 39-42, document vectors are arranged into clusters) comprising: clustering first documents in a first dataset to produce first document classes; (in col. 2, lines 39-42, document vectors are arranged into clusters), and creating centroid seeds based on said first document classes (in col. 2, lines 43-45, the invention finds centroids). However, Lantrip fails to disclose clustering second documents in a second dataset using said centroid seeds. However, in col. 14, lines 10-45 of Ruocco, Ruocco discloses in the claim processing in parallel second datasets based on cluster information

from previous cluster vectors (see col. 14, lines 28-30) in order to gain the benefit of information from previous clusters to improve analysis of subsequent datasets. Ruocco's invention further may be interpreted such that said second dataset has a similar clustering to that of said first dataset (as the term "similar" is sufficiently broad that any two given datasets would have some degree of similarity, see35 U.S.C. 112 rejection, above.), further wherein said second dataset comprises a new, but related dataset different than said first dataset (once the first dataset is transformed, it is by definition a new, but related dataset). It would have been obvious to one of ordinary skill in the art at the time of the invention to use the information contained in the centroid seeds from Lantrip for subsequent datasets as in Ruocco in order to improve analysis of subsequent datasets.

Regarding dependent claim 11, Lantrip fails to disclose a method further comprising generating a second vector space model by counting, for each word in said first dictionary, a number of said second document in which said word occurs. However, Ruocco, in col. 14, lines 20-35, discloses generating such a vector space model for multiple document sets in order to aid in the clustering analysis of the document sets. It would have been obvious to one of ordinary skill in the art at the time of the invention to generate a second vector space model in the manner of Ruocco in Lantrip's invention in order to aid in the clustering analysis of the document sets.

Regarding dependent claim 12, Lantrip discloses that said creating of said centroid seeds comprises: classifying said second vector space model using said first document classes to produce a classified second vector space model (col. 2, lines 39- 42, the vector space model is clustered); and determining a mean of vectors in each class in said classified second vector space model, wherein said mean comprises said centroid seeds (col. 2, lines 43-45, the centroid is the center of mass of the clusters).

Regarding dependent claim 13, Lantrip and Ruocco fail to disclose a method further comprising forming a second dictionary of most common words in said second dataset; generating a third vector space model by counting, for each word in said second dictionary, a number of said second documents in which said word occurs; and clustering said documents in said second dataset based on said third vector space model to produce

a second dataset cluster. However, this constitutes simply extending and repeating claim 3 to a third dataset, and it was notoriously well known in the art at the time of the invention that it is useful to repeat steps for multiple datasets to take advantage of their utility for subsequent data. It would have been obvious to one of ordinary skill in the art at the time of the invention to extend the steps of claim 3 to a subsequent dataset to gain the benefits of the analysis for that dataset.

Regarding dependent claim 14, Lantrip discloses, in col. 2, lines 39-45 that clustering of said documents in said dataset using said centroid seeds produces an adapted dataset cluster. However, Lantrip fails to disclose the use of multiple datasets and that the method further comprises comparing classes in said adapted dataset cluster to classes in said second dataset cluster; and adding classes to said adapted dataset cluster based on said comparing. However, in col. 4, lines 61-67, Rocco deals with comparing multiple dataset clusters in order to obtain more information about the relative status of the datasets. It would have been obvious to one of ordinary skill in the art at the time of the invention to compare multiple dataset clusters in order to obtain more information about the relative status of the datasets.

Regarding independent claim 15, it is essentially analogous to claim 1 except that it involves the steps of generating a vector space model of said second documents, which Ruocco presents in col. 14, lines 27-36, and classifying said vector space model of said second documents using said first document classes to produce a classified vector space model, which Ruocco presents in col. 14, lines 27-36. It would have been obvious to one of ordinary skill in the art at the time of the invention to use the Ruocco form of vector space analysis in addition to the Lantrip material from the rejection of Claim 1 in order to enhance the classifications of the two datasets. The result would produce an invention that would serve to reject claim 15:

Regarding dependent claim 17, the applicant discloses the limitations substantially similar to those in claim 11. Claim 17 is similarly rejected.

Regarding dependent claim 18, the applicant discloses the limitations substantially similar to those in claim 13. Claim 18 is similarly rejected.

Regarding dependent claim 19, the applicant discloses the limitations substantially similar to those in claim 14. Claim 18 is similarly rejected.

Regarding independent claim 20, Lantrip discloses a method of clustering documents comprising: forming a first dictionary of most common words in a first dataset (col. 2, lines 30-35, Lantrip forms a first dictionary of common words); generating a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said words occurs (col. 2, lines 35-40, Lantrip forms vectors); clustering said first documents in said first dataset based on said first vector space model to produce first document classes (col. 2, lines 39-42, Lantrip forms clusters), and determining a mean of vectors in each class in said classified second vector space model to produce centroid seeds; (col. 2, lines 43-45, Lantrip forms centroid seeds) and clustering documents in a second datasets using said centroid seeds (col. 2, lines 45-57, Lantrip clusters using centroids). Lantrip fails to disclose generating a second vector space model by counting, for each word in said first dictionary, and number of said second documents in which said word occurs and classifying said second documents in said second vector space model using said first document classes to produce a classified second vector space model. However, col. 14, lines 28-36 of Ruocco indicate that vector clustering analysis may involve multiple datasets in order to gain the benefit of information analysis from multiple sources. It would have been obvious to one of ordinary skill in the art at the time of the invention to have vector clustering analysis involve multiple datasets in order to gain the benefit of information analysis from multiple sources.

Regarding dependent claim 21, the applicant discloses the limitations substantially similar to those in claim 13. Claim 21 is similarly rejected.

Regarding dependent claim 22, the applicant discloses the limitations substantially similar to those in claim 14. Claim 22 is similarly rejected.

Regarding independent claim 23, , the applicant discloses the limitations substantially similar to those in claim 8. Claim 23 is similarly rejected.

Regarding dependent claim 26, the applicant discloses the limitations substantially similar to those in claim 11. Claim 26 is similarly rejected.

Regarding dependent claim 27, the applicant discloses the limitations substantially similar to those in claim 12. Claim 27 is similarly rejected.

Regarding dependent claim 28, the applicant discloses the limitations substantially similar to those in claim 13. Claim 28 is similarly rejected.

Regarding dependent claim 29, the applicant discloses the limitations substantially similar to those in claim 14. Claim 29 is similarly rejected.

Applicant's arguments filed 1 November 2006 have been fully considered but they are not persuasive.

The applicant argues that Ryocco fails to teach that the second data set has a similar, based on said centroid seeds, clustering to that of said first dataset (page 10). The examiner respectfully disagrees. Ryocco suggests clustering second documents in a second dataset using said centroid seeds, such that said second dataset has a similar clustering to that of said first dataset (pages 9-10). Although, the applicant argues that Ryocco uses the centroid seeds of the first dataset, the claim limitations require "using said centroid seeds (claim 8, line 15; emphasis added)." Although these centroid seeds may be used with the first document, the applicant's plain claim languag~'e restricts using a second set of centroid seeds, and instead requires the original centroid seeds be used).

B. Appellants' Position

1. Independent Claim 8.

The Applicants submit that a prima facie case of obvious has not been established in support of rejecting claim 8 (See *In re Vaeck*, 947 F.2d 488, 20 USPQ2d 1438 (Fed. Cir. 1991)). Specifically, the cited prior art references as a whole do not suggest the desirability and, thus, the obviousness of making the combination (see *Hodosh v. Block*

Drug Co., Inc., 786 F.2d 1136, 1143 n.5, 229 USPQ 182, 187 n.5 (Fed. Cir. 1986)). Additionally, the combination of Ruocco and Lantrip does not teach or suggest all of the limitations of claim 8.

(a) Lack Of Suggestion/Motivation To Combine.

The Applicants submit that the there is no suggestion or motivation, either in the references themselves or in the knowledge generally available to one of ordinary skill in the art, to combine the teachings of Lantrip and Ruocco.

Lantrip discloses a method of visually displaying the relative content of a large number of documents (see Abstract). Specifically, the relationships of the documents are presented in a 3D landscape with the relative sizes and heights of peaks in the landscape representing the relative significance of a relationship of an individual document in the document set to a topic or term (see col. 2, lines 26-31). As discussed in col. 2, lines 30-55, the Lantrip method includes the steps of creating a vector for each document in the set such that each vector represents the relative relationship of an individual document to a term or topic. Then, the vectors are arranged into clusters. For each cluster, the "centroid coordinates" (i.e., the coordinates for the center of the cluster mass) are determined as well as the distance of each document in a cluster from the centroid. This information is ultimately used to generate the 3D display. Thus, Lantrip is only concerned with a single set of documents.

Ruocco discloses a processing system that utilizes parallel processors for organizing and clustering a large number of documents (see Abstract). Col. 4, lines 35-45, of Ruocco summarizes conventional document clustering in which text documents are converted to document vectors and clustered. Clusters of documents are represented by cluster vectors. As each new document is processed, its document vector is compared to the cluster vector for each cluster. If a document vector is similar to a given cluster vector, it is placed in that cluster. If it is not, a new cluster is formed. Col. 4, lines 45-60, summarizes the Ruocco method that uses multiple processors to store various clusters. In the Ruocco method, as each new document in a dataset is processed, its document vector is simultaneously compared to P cluster vectors, where P is the number of processors

within the system. That is, the processors work simultaneously and each processor compares the document vector to each of its clusters.

The Office Action indicates that it "would have been obvious to one of ordinary skill in the art at the time of the invention to use the information contained in the centroid seeds from Lantrip for subsequent datasets as in Ruocco in order to improve analysis of subsequent datasets." However, as discussed above, Lantrip does not disclose "centroid seeds", but rather calculating centroid coordinates for clusters of documents in a single dataset in order to generate a 3D display of that dataset. Rocco on the other hand deals with a single dataset of documents and clustering those documents on different processors. Neither address how different, but related datasets, would be processed.

Specifically, the starting points for creating clusters in a dataset are typically selected in some random manner. In Ruocco, the starting point is based on the first document vector and the document vectors of any subsequently processed document that is not similar to already established clusters. In Landrip, the starting point is not disclosed. In the present invention, however, it is recognized that clustering documents from new and different, but related datasets, results in a lack of continuity that is a drawback when one in interested in tracking changes in the data (trends) over time (see page 5, lines 0-9). This is a problem not addressed in either Ruocco or Landrip (as they disclose conventional techniques for selecting the starting points and no technique, respectively). The present invention solves this problem by intentionally biasing the clustering algorithm towards first classes in a first dataset by using centroid seeds generated from clusters corresponding to those first classes as the starting points for clustering a second new and different, but related, dataset. More specifically, the claimed invention solves the problem of finding new categories in a second data set that did not exist in the first data set, while at the same time maintaining as nearly as possible categories from the first data set as categories in the second data set. With the claimed invention, there is no requirement, and in fact it is not assumed, that the first and second datasets have any of the same data elements in them. They are allowed to have some of the same elements, but this is in no way a requirement for the claimed invention. The

claimed invention is designed in such a way as to find the similarities between the two datasets, where they exist, while at the same time finding the key differences (emerging concepts) in the second dataset.

Nothing in either Ruocco or Lantrip suggests that subsequent datasets (related or not) would be processed any differently then stated in either reference, much less that they would be processed based on information from prior datasets (as in the present invention). Furthermore, nothing in Lantrip or Ruocco suggests that the centroid coordinates used to make the 3D display of Lantrip would be used for anything other than forming a 3D display, much less that they would be used as the seeds for forming the initial clusters for subsequent datasets to be processed according to the Ruocco invention (i.e., on multiple processors). Rather the only obvious Lantrip and Ruocco combination would result in the calculation of centroid coordinates for all of the clusters stored on all of the processors in of Ruocco in order to generate a 3D display as in Lantrip.

(b) All Of The Claim Limitations Not Taught.

The Applicants further submit that neither Lantrip, nor Ruocco, teach or suggest the following features of the system of independent claim 8: (1) "a centroid seed generator operative to generate centroid seeds based on said first document classes"; (2) "a dictionary generator adapted to generate a first dictionary of most common words in said first dataset;" (3) "a vector space model generator adapted to generate a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said word occurs"; (4) "wherein said cluster generator clusters second documents in said second dataset using said centroid seeds such that said second dataset has a similar, based on said centroid seeds, clustering to that of said first dataset"; and (5) "wherein said second dataset comprises a new, but related, based on said centroid seeds, dataset different than said first dataset."

More specifically, the Office Action did not address the claimed invention of a "system for clustering documents in datasets", nor did it address each of the claimed features thereof. That is, in rejecting claim 8, the Office Action discussed a method of

clustering documents not a system (as claimed) and further addressed several method steps but not the structural features of the system (as claimed). Additionally, the method steps addressed in the Office Action included some steps which might reasonably be performed by the claimed structural features, but excluded others. For example, the Office Action did not address the feature of a "centroid seed generator", but did indicate a method step of "creating centroid seeds" was taught by Lantrip. The Office Action also did not address the features of "a dictionary generator" or "a vector space model generator". Furthermore, it did not indicate any method steps taught by Lantrip or Ruocco during which such a first dictionary or first vector space model would be generated.

As to the features that were addressed, the Office Action provides that Lantrip discloses "creating centroid seeds based on said first document classes (in col. 2, lines 43-45, the invention finds centroids)." The Applicants respectfully disagree. Lantrip does not disclose creating centroid seeds, as indicated in the Office Action, but rather col. 2, lines 43-45 of Lantrip refers to calculating "centroid coordinates" of clusters in a dataset (i.e., determining the coordinates for the center mass of each cluster after all the clusters are formed). Thus, Lantrip discloses centroid coordinates, but not "centroid seeds." The centroid coordinates of Lantrip are mapped points in a cluster that are used by Lantrip to ultimately generate a 3D display for visually representing the dataset. They are not "seeds" that are subsequently used as the starting points to create clusters for a new and different, but related, dataset.

As explained in pages 11, lines 5-20 of the specification of the present invention, each "centroid seed" is an average value of all the values in a document class (i.e., as defined by a cluster) from one dataset and each of these seeds are subsequently used to generate an initial cluster in another dataset. That is, conventionally, the starting points for creating clusters in a dataset are selected in some random manner. However, the present invention intentionally biases the clustering algorithm towards the classes in a first dataset by using centroid seeds generated from those first classes as the starting points for the clusters of a second new and different, but related, dataset.

The Office Action further acknowledges that Lantrip does not disclose "clustering second documents in a second dataset using said centroid seeds." Therefore, the Office Action provides that "in col. 14, lines 10-45 of Ruocco, Ruocco discloses in the claim processing in parallel second datasets based on cluster information from previous cluster vectors (see col. 14, lines 28-30) in order to gain the benefit of information from previous clusters to improve analysis of datasets. Ruocco's invention further may be interpreted such that said second dataset has a similar clustering to that of said first dataset (as the term "similar" is sufficiently broad that any two given datasets would have some degree of similarity, see 35 U.S.C. 112 rejection, above.), further wherein said second data set comprises a new, but related dataset different than said first dataset (once the first dataset is transformed, it is by definition, a new, but related dataset). It would have been obvious to one of ordinary skill in the art at the time of the invention to use the information contained in the centroid seeds from Lantrip for subsequent datasets as in Ruocco in order to improve analysis of subsequent datasets." The Applicants respectfully disagree.

Ruocco only discloses clustering documents in a single dataset. It does not teach or suggest clustering first documents in a first dataset and then subsequently clustering second documents in a second dataset using centroid seeds generated based on the first document clusters (i.e., based on the first document classes). Specifically, the cited portion of Ruocco discloses a computer system having parallel processors and a method which individually examines documents and assigns similar documents to a particular cluster that is in turn assigned to a particular processor. The method is initiated by selecting a first document, converting it into a first document vector, designating the first document vector as a first cluster vector and assigning the first cluster vector to a first processor (see col. 14, lines 18-27). Then, a second document is selected and converted into a second document vector. This second document vector is compared to the first document vector. If they are similar the second document vector is assigned to the first cluster in the first processor. If they are not, the second document vector is assigned to a second processor (see col. 14, lines 27-35).

Thus, Ruocco simply discloses an improvement over legacy methods of clustering documents in a single dataset (e.g., a 32 document set, a 64 document set, a 128 document set (see col. 7, lines 30-35). That is, due to a determined need to organize a large number of documents in a document set, the Ruocco method forms clusters of similar documents, but instead of maintaining those clusters on a single processor, Ruocco maintains one or more clusters on multiple different processors. The initial cluster vector for a given cluster in Ruocco is disclosed as being the first document vector in the cluster. After that it is a mathematical average of document vectors in the cluster (see col. 14, line 25-26). As each new document is added to the dataset, the document is assigned to a particular cluster on a particular processor based on similarities between its document vector and the cluster vector of the particular cluster (see col. 14, lines 28-46).

Nowhere in the cited portion of Ruocco does it teach or disclose "processing in parallel second datasets based on cluster information from previous cluster vectors in order to gain the benefit of information from previous clusters to improve analysis of datasets," as indicated by the Office Action. While Ruocco uses cluster vectors (i.e., either the document vector if there is only one document in a cluster or the mathematical average of document vectors in a cluster of more than one document) when determining whether or not to add a new document to an old cluster or to create a new cluster in a dataset, Ruocco does not teach or suggest using any information from the first document classes (i.e., first clusters) of a first dataset when clustering second documents in a second dataset. That is, presumably, for each new dataset (i.e., for each new set of documents) in Ruocco, the method would follow the exact same steps as disclosed in the claim set in column 14 such that the first document vector will be designated as the first cluster vector and so on.

Therefore, the Applicants submit that the applied prior art references alone and in combination do teach or suggest the following features of independent claim 8: (1) "a centroid seed generator operative to generate centroid seeds based on said first document classes"; (2) "a dictionary generator adapted to generate a first dictionary of most common words in said first dataset;" (3) "a vector space model generator adapted to

generate a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said word occurs"; (4) "wherein said cluster generator clusters second documents in said second dataset using said centroid seeds such that said second dataset has a similar, based on said centroid seeds, clustering to that of said first dataset"; and (5) "wherein said second dataset comprises a new, but related, based on said centroid seeds, dataset different than said first dataset." Further, dependent claims 11-14 are similarly patentable, not only by virtue of their dependency from a patentable claim, but also by virtue of the additional features of the invention they define. In view of the foregoing, the Examiner is respectfully requested to reconsider and withdraw this rejection.

2. Independent Claim 15.

The Applicants submit that a prima facie case of obvious has not been established in support of rejecting claim 15 (See *In re Vaeck*, 947 F.2d 488, 20 USPQ2d 1438 (Fed. Cir. 1991)). Specifically, the cited prior art references as a whole do not suggest the desirability and, thus, the obviousness of making the combination (see *Hodosh v. Block Drug Co., Inc.*, 786 F.2d 1136, 1143 n.5, 229 USPQ 182, 187 n.5 (Fed. Cir. 1986)). Additionally, the combination of Ruocco and Lantrip does not teach or suggest all of the limitations of claim 15.

(a) Lack Of Suggestion/Motivation To Combine.

The Applicants submit that the there is no suggestion or motivation, either in the references themselves or in the knowledge generally available to one of ordinary skill in the art, to combine the teachings of Lantrip and Ruocco.

Lantrip discloses a method of visually displaying the relative content of a large number of documents (see Abstract). Specifically, the relationships of the documents are presented in a 3D landscape with the relative sizes and heights of peaks in the landscape representing the relative significance of a relationship of an individual document in the document set to a topic or term (see col. 2, lines 26-31). As discussed in

col. 2, lines 30-55, the Lantrip method includes the steps of creating a vector for each document in the set such that each vector represents the relative relationship of an individual document to a term or topic. Then, the vectors are arranged into clusters. For each cluster, the "centroid coordinates" (i.e., the coordinates for the center of the cluster mass) are determined as well as the distance of each document in a cluster from the centroid. This information is ultimately used to generate the 3D display. Thus, Lantrip is only concerned with a single set of documents.

Ruocco discloses a processing system that utilizes parallel processors for organizing and clustering a large number of documents (see Abstract). Col. 4, lines 35-45, of Ruocco summarizes conventional document clustering in which text documents are converted to document vectors and clustered. Clusters of documents are represented by cluster vectors. As each new document is processed, its document vector is compared to the cluster vector for each cluster. If a document vector is similar to a given cluster vector, it is placed in that cluster. If it is not, a new cluster is formed. Col. 4, lines 45-60, summarizes the Ruocco method that uses multiple processors to store various clusters. In the Ruocco method, as each new document in a dataset is processed, its document vector is simultaneously compared to P cluster vectors, where P is the number of processors within the system. That is, the processors work simultaneously and each processor compares the document vector to each of its clusters.

The Office Action indicates that it "would have been obvious to one of ordinary skill in the art at the time of the invention to use the information contained in the centroid seeds from Lantrip for subsequent datasets as in Ruocco in order to improve analysis of subsequent datasets." However, as discussed above, Lantrip does not disclose "centroid seeds", but rather calculating centroid coordinates for clusters of documents in a single dataset in order to generate a 3D display of that dataset. Rocco on the other hand deals with a single dataset of documents and clustering those documents on different processors. Neither address how different, but related datasets, would be processed.

Specifically, the starting points for creating clusters in a dataset are typically selected in some random manner. In Ruocco, the starting point is based on the first

document vector and the document vectors of any subsequently processed document that is not similar to already established clusters. In Landrip, the starting point is not disclosed. In the present invention, however, it is recognized that clustering documents from new and different, but related datasets, results in a lack of continuity that is a drawback when one in interested in tracking changes in the data (trends) over time (see page 5, lines 0-9). This is a problem not addressed in either Ruocco or Landrip (as they disclose conventional techniques for selecting the starting points and no technique, respectively). The present invention solves this problem by intentionally biasing the clustering algorithm towards first classes in a first dataset by using centroid seeds generated from clusters corresponding to those first classes as the starting points for clustering a second new and different, but related, dataset. More specifically, the claimed invention solves the problem of finding new categories in a second data set that did not exist in the first data set, while at the same time maintaining as nearly as possible categories from the first data set as categories in the second data set. With the claimed invention, there is no requirement, and in fact it is not assumed, that the first and second datasets have any of the same data elements in them. They are allowed to have some of the same elements, but this is in no way a requirement for the claimed invention. The claimed invention is designed in such a way as to find the similarities between the two datasets, where they exist, while at the same time finding the key differences (emerging concepts) in the second dataset.

Nothing in either Ruocco or Lantrip suggests that subsequent datasets (related or not) would be processed any differently then stated in either reference, much less that they would be processed based on information from prior datasets (as in the present invention). Furthermore, nothing in Lantrip or Ruocco suggests that the centroid coordinates used to make the 3D display of Lantrip would be used for anything other than forming a 3D display, much less that they would be used as the seeds for forming the initial clusters for subsequent datasets to be processed according to the Ruocco invention (i.e., on multiple processors). Rather the only obvious Lantrip and Ruocco combination

would result in the calculation of centroid coordinates for all of the clusters stored on all of the processors in of Ruocco in order to generate a 3D display as in Lantrip.

(b) All Of The Claim Limitations Not Taught.

The Applicants further submit that neither Lantrip, nor Ruocco, teach or suggest the following features of the method of independent claim 15: (1) "generating a vector space model of said second documents"; (2) "classifying said vector space model of said second documents using said first document classes to produce a classified vector space model"; (3) "determining a mean of vectors in each class in said classified vector space model to produce centroid seeds"; (4) "clustering said second documents using said centroid seeds, such that said second dataset has a similar, based on said centroid seeds, clustering to that of said first dataset"; (5) "wherein said second dataset comprises a new, but related, based on said centroid seeds, dataset different than said first dataset"; and (5) "wherein … said clustering of said first documents in said first data comprises: forming a first dictionary of most common words in said first dataset; and generating a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said word occurs."

More specifically, in rejecting claim 15, the Office Action provides that, "it is essentially analogous to claim 1 except that it involves the steps of generating a vector space model of said second documents, which Ruocco present in col. 14, lines 27-36, and classifying said vector space model of said second documents using said first document classes to produce a classified vector space model, which Ruocco presents in col. 14, lines 27-36. It would have been obvious to one of ordinary skill in the art at the time of the invention to use the Ruocco form of vector space analysis in addition to the Lantrip material from the rejection of claim 1 in order to enhance the classifications of the two datasets." It should be noted that since claim 1 is and was not pending at the time of the final office action. Therefore, it is assumed that the Examiner was referring to the explanation provided for rejecting claim 8 rather than claim 1.

Claim 15 includes the features of "generating a vector space model of said second documents" and "classifying said vector space model of said second document using said first document classes to produce a classified vector space model". The Office Action cites col. 14, lines 27-36 of Ruocco as teaching these features. The Applicants respectfully disagree. Col. 14, lines 27-36 refers to a feature in claim 1 of Ruocco, namely, "selecting a second electronic document and comparing the vector of the second electronic document with the first cluster vector to determine if the second document has similar characteristics, and assigning the second document vector to the first cluster vector if they have similar characteristics or designating the second document vector as a second cluster vector and assigning the second cluster vector to a second processor of the parallel processors if there are different characteristics." The Applicants submit that this feature does not equate to the claimed features of "generating a vector space model of said second documents" and "classifying said vector space model of said second documents using said first document classes to produce a classified vector space model". Rather this feature refers to that aspect of the Ruocco invention which uses cluster vectors (i.e., either the document vector if there is only one document in a cluster or the mathematical average of document vectors in a cluster of more than one document) when determining whether or not to add a new document to an old cluster or to create a new cluster in a dataset.

Specifically, as discussed above, Ruocco discloses an improvement over legacy methods of clustering documents in a single dataset (e.g., a 32 document set, a 64 document set, a 128 document set (see col. 7, lines 30-35). Due to a determined need to organize a large number of documents in a document set, the Ruocco method forms clusters of similar documents, but instead of maintaining those clusters on a single processor, Ruocco maintains one or more clusters on multiple different processors. The initial cluster vector for a given cluster in Ruocco is disclosed as being the first document vector in the cluster. After that it is a mathematical average of document vectors in the cluster (see col. 14, line 25-26). As each new document is added to the dataset, the document is assigned to a particular cluster on a particular processor based on similarities

between its document vector and the cluster vector of the particular cluster (see col. 14, lines 28-46). Thus, Ruocco uses cluster vectors (i.e., either the document vector if there is only one document in a cluster or the mathematical average of document vectors in a cluster of more than one document) when determining whether or not to add a new document to an old cluster or to create a new cluster in a dataset. However, nowhere in Ruocco does not it teach or suggest "generating a vector space model of said second documents [in a second dataset]" and "classifying said vector space model of said second documents [in the second dataset] using said first document classes [of the first documents in the first dataset] to produce a classified vector space model."

Claim 15 also includes the feature of "determining a mean of vectors in each class in said classified vector space model to produce centroid seeds". Neither this feature, nor any similar feature was claimed in claim 8. The Office Action does not discuss this feature (either indirectly by reference to claim 8 or directly) and the Applicants submit that nowhere in the cited prior art references is this feature taught or suggested. It should be noted that in rejecting claim 8, the Office Action provided that Lantrip disclosed "creating centroid seeds". As discussed above, the Applicants respectfully disagree. Col. 2, lines 43-45 of Lantrip refers to calculating "centroid coordinates" of clusters in a dataset (i.e., determining the coordinates for the center mass of each cluster after all the clusters are formed). These centroid coordinates of Lantrip are mapped points in a cluster that are used by Lantrip to ultimately generate a 3D display representing the dataset. They are not "centroid seeds" that are subsequently used as the starting points to create clusters for a new and different, but related, dataset. Furthermore, the "centroid coordinates" of Lantrip were not produced by determining a mean of vectors in each class in the classified vector space model.

Claim 15 also includes the feature of "clustering said second documents using said centroid seeds, such that said second dataset has a similar, based on said centroid seeds, clustering to that of said first dataset". The Office Action did not address the feature directly but rather indirectly by reference to claim 8. Specifically, in rejecting claim 8, the Office Action stated that Lantrip "failed to disclose clustering second

documents in a second dataset using said centroid seeds. However, in col. 14, lines 10-45 of Ruocco, Ruocco discloses in the claim processing in parallel second datasets based on cluster information from previous cluster vectors (see col. 14, lines 28-30) in order to gain the benefit of information from previous clusters to improve analysis of subsequent datasets. Ruocco's invention further may be interpreted such that said second dataset has a similar clustering to that of said first dataset (as the term "similar" is sufficiently broad that any two given datasets would have some degree of similarity, see35 U.S.C. 112 rejection, above.), further wherein said second dataset comprises a new, but related dataset different than said first dataset (once the first dataset is transformed, it is by definition a new, but related dataset). It would have been obvious to one of ordinary skill in the art at the time of the invention to use the information contained in the centroid seeds from Lantrip for subsequent datasets as in Ruocco in order to improve analysis of subsequent datasets." The Applicants respectfully disagree.

Ruocco uses cluster vectors (i.e., either the document vector if there is only one document in a cluster or the mathematical average of document vectors in a cluster of more than one document) when determining whether or not to add a new document to an old cluster or to create a new cluster in a dataset. The centroid coordinates of Lantrip are used for mapping 3D displays. As discussed above, the Applicants submit that nothing in either Ruocco or Lantrip suggest that subsequent datasets (related or not) would or should be processed any different. Consequently, it would not have been obvious to use the coordinates generated for a 3D display of one dataset processed according to the Ruocco technique as the starting points (i.e., centroid seeds) for clusters of a new dataset to be processed according to that Ruocco technique.

Claim 15 also includes the feature of "wherein … said clustering of said first documents in said first data comprises: forming a first dictionary of most common words in said first dataset; and generating a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said word occurs." The Office Action does not discuss this feature (either indirectly by reference to claim 8 or

directly) and the Applicants submit that nowhere in the cited prior art references is this feature taught or suggested.

Therefore, the Applicants submit that the applied prior art references alone and in combination do teach or suggest the following features of independent claim 15: (1) "generating a vector space model of said second documents"; (2) "classifying said vector space model of said second documents using said first document classes to produce a classified vector space model"; (3) "determining a mean of vectors in each class in said classified vector space model to produce centroid seeds"; (4) "clustering said second documents using said centroid seeds, such that said second dataset has a similar, based on said centroid seeds, clustering to that of said first dataset"; (5) "wherein said second dataset comprises a new, but related, based on said centroid seeds, dataset different than said first dataset"; and (5) "wherein … said clustering of said first documents in said first data comprises: forming a first dictionary of most common words in said first dataset; and generating a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said word occurs." Furthermore, dependent claims 17-19 are similarly patentable, not only by virtue of their dependency from a patentable claim, but also by virtue of the additional features of the invention they define. In view of the foregoing, the Examiner is respectfully requested to reconsider and withdraw this rejection.

3. Independent Claim 20.

The Applicants submit that a prima facie case of obvious has not been established in support of rejecting claim 20 (See *In re Vaeck*, 947 F.2d 488, 20 USPQ2d 1438 (Fed. Cir. 1991)). Specifically, the cited prior art references as a whole do not suggest the desirability and, thus, the obviousness of making the combination (see *Hodosh v. Block Drug Co., Inc.*, 786 F.2d 1136, 1143 n.5, 229 USPQ 182, 187 n.5 (Fed. Cir. 1986)). Additionally, the combination of Ruocco and Lantrip does not teach or suggest all of the limitations of claim 20.

a. Lack Of Suggestion/Motivation To Combine.

The Applicants submit that the there is no suggestion or motivation, either in the references themselves or in the knowledge generally available to one of ordinary skill in the art, to combine the teachings of Lantrip and Ruocco.

Lantrip discloses a method of visually displaying the relative content of a large number of documents (see Abstract). Specifically, the relationships of the documents are presented in a 3D landscape with the relative sizes and heights of peaks in the landscape representing the relative significance of a relationship of an individual document in the document set to a topic or term (see col. 2, lines 26-31). As discussed in col. 2, lines 30-55, the Lantrip method includes the steps of creating a vector for each document in the set such that each vector represents the relative relationship of an individual document to a term or topic. Then, the vectors are arranged into clusters. For each cluster, the "centroid coordinates" (i.e., the coordinates for the center of the cluster mass) are determined as well as the distance of each document in a cluster from the centroid. This information is ultimately used to generate the 3D display. Thus, Lantrip is only concerned with a single set of documents.

Ruocco discloses a processing system that utilizes parallel processors for organizing and clustering a large number of documents (see Abstract). Col. 4, lines 35-45, of Ruocco summarizes conventional document clustering in which text documents are converted to document vectors and clustered. Clusters of documents are represented by cluster vectors. As each new document is processed, its document vector is compared to the cluster vector for each cluster. If a document vector is similar to a given cluster vector, it is placed in that cluster. If it is not, a new cluster is formed. Col. 4, lines 45-60, summarizes the Ruocco method that uses multiple processors to store various clusters. In the Ruocco method, as each new document in a dataset is processed, its document vector is simultaneously compared to P cluster vectors, where P is the number of processors within the system. That is, the processors work simultaneously and each processor compares the document vector to each of its clusters.

The Office Action indicates that it "would have been obvious to one of ordinary skill in the art at the time of the invention to use the information contained in the centroid

seeds from Lantrip for subsequent datasets as in Ruocco in order to improve analysis of subsequent datasets." However, as discussed above, Lantrip does not disclose "centroid seeds", but rather calculating centroid coordinates for clusters of documents in a single dataset in order to generate a 3D display of that dataset. Rocco on the other hand deals with a single dataset of documents and clustering those documents on different processors. Neither address how different, but related datasets, would be processed.

Specifically, the starting points for creating clusters in a dataset are typically selected in some random manner. In Ruocco, the starting point is based on the first document vector and the document vectors of any subsequently processed document that is not similar to already established clusters. In Landrip, the starting point is not disclosed. In the present invention, however, it is recognized that clustering documents from new and different, but related datasets, results in a lack of continuity that is a drawback when one in interested in tracking changes in the data (trends) over time (see page 5, lines 0-9). This is a problem not addressed in either Ruocco or Landrip (as they disclose conventional techniques for selecting the starting points and no technique, respectively). The present invention solves this problem by intentionally biasing the clustering algorithm towards first classes in a first dataset by using centroid seeds generated from clusters corresponding to those first classes as the starting points for clustering a second new and different, but related, dataset. More specifically, the claimed invention solves the problem of finding new categories in a second data set that did not exist in the first data set, while at the same time maintaining as nearly as possible categories from the first data set as categories in the second data set. With the claimed invention, there is no requirement, and in fact it is not assumed, that the first and second datasets have any of the same data elements in them. They are allowed to have some of the same elements, but this is in no way a requirement for the claimed invention. The claimed invention is designed in such a way as to find the similarities between the two datasets, where they exist, while at the same time finding the key differences (emerging concepts) in the second dataset.

Nothing in either Ruocco or Lantrip suggests that subsequent datasets (related or not) would be processed any differently then stated in either reference, much less that they would be processed based on information from prior datasets (as in the present invention). Furthermore, nothing in Lantrip or Ruocco suggests that the centroid coordinates used to make the 3D display of Lantrip would be used for anything other than forming a 3D display, much less that they would be used as the seeds for forming the initial clusters for subsequent datasets to be processed according to the Ruocco invention (i.e., on multiple processors). Rather the only obvious Lantrip and Ruocco combination would result in the calculation of centroid coordinates for all of the clusters stored on all of the processors in of Ruocco in order to generate a 3D display as in Lantrip.

b. All Of The Claim Limitations Not Taught.

The Applicants further submit that neither Lantrip, nor Ruocco, teach or suggest the following features of the method of independent claim 20: (1) "forming a first dictionary of most common words in a first dataset"; (2) "generating a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said word occurs"; (3) "clustering said first documents in said first dataset based on said first vector space model to produce first document classes"; (4) "generating a second vector space model by counting, for each word in said first dictionary, a number of said second documents in which said word occurs"; (5) "classifying said second documents in said second vector space model using said first document classes to produce a classified second vector space model"; (6) "determining a mean of vectors in each class in said classified second vector space model to produce centroid seeds"; (7) "clustering second documents in a second dataset using said centroid seeds, such that said second dataset has a similar, based on said centroid seeds, clustering to that of said first dataset"; and (8)"wherein said second dataset comprises a new, but related ,based on said centroid seeds, dataset different than said first dataset."

The Office Action cites col. 2, lines 30-35 of Lantrip as disclosing "forming a first dictionary of most common words in a first dataset." The Applicants respectfully

disagree. Lantrip col. 2, lines 27-36, provides: "The relationships of a plurality of documents are presented in a three-dimensional landscape with the relative size and height of a peak in the three-dimensional landscape representing the relative significance of the relationship of a topic, or term, and the individual document in the document set. The steps of the process are: (a) constructing an electronic database of a plurality of documents to be analyzed; b) creating a plurality of high dimensional vectors, one for each of the plurality of documents, such that each of the high dimensional vectors represents the relative relationship of the individual documents to the term, or topic attribute;". Nothing in the cited portion of Lantrip teaches or suggests the feature of "forming a first dictionary of most common words in a first dataset."

The Office Action provides that Lantrip discloses "generating a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said words occurs (col. 2, lines 35-40, Lantrip forms vectors)". The Applicants respectfully disagree. Lantrip, col. 2, lines 35-42 provides: "(b) creating a plurality of high dimensional vectors, one for each of the plurality of documents, such that each of the high dimensional vectors represents the relative relationship of the individual documents to the term, or topic attribute; (c) arranging the high dimensional vectors into clusters, with each of the clusters representing a plurality of documents grouped by relative significance of their relationship to a topic attribute". While Lantrip does create vectors and arrange them into clusters, nothing in Lantrip discloses the claimed feature of "generating a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said words occurs".

The Office Action provides that Lantrip discloses "clustering said first documents in said first dataset based on said first vector space model to produce first document classes (col. 2, lines 39-42, Lantrip forms clusters),". The Applicants respectfully disagree. The Applicants respectfully disagree. As discussed above, Lantrip, col. 2, lines 35-42 provides: "(b) creating a plurality of high dimensional vectors, one for each of the plurality of documents, such that each of the high dimensional vectors represents the relative relationship of the individual documents to the term, or topic attribute; (c)

arranging the high dimensional vectors into clusters, with each of the clusters representing a plurality of documents grouped by relative significance of their relationship to a topic attribute". Again, while Lantrip does create vectors and arrange them into clusters, nothing in Lantrip discloses the claimed feature of "clustering said first documents in said first dataset based on said first vector space model to produce first document classes."

The Office Action provides that Lantrip discloses "determining a mean of vectors in each class in said classified second vector space model to produce centroid seeds; (col. 2, lines 43-45, Lantrip forms centroid seeds)". The Applicants respectfully disagree. As discussed above, the Applicants respectfully disagree. Col. 2, lines 43-45 of Lantrip refers to calculating "centroid coordinates" of clusters in a dataset (i.e., determining the coordinates for the center mass of each cluster after all the clusters are formed). These centroid coordinates of Lantrip are mapped points in a cluster that are used by Lantrip to ultimately generate a 3D display representing the dataset. They are not "centroid seeds" that are subsequently used as the starting points to create clusters for a new and different, but related, dataset. Furthermore, the "centroid coordinates" of Lantrip were not produced by determining a mean of vectors in each class in the classified vector space model.

The Office Action provides that Lantrip discloses: "clustering documents in a second datasets using said centroid seeds (col. 2, lines 45-57, Lantrip clusters using centroids)." The Applicants respectfully disagree. Lantrip, col. 2, lines 45-57, provides: "(e) constructing a vector for each document, with each vector containing the distance from the document to each centroid coordinate in high-dimensional space; (f) creating a plurality of term (or topic) layers, each of the term layers corresponding to a descriptive term (or topic) applied to each cluster, and identifying x,y coordinates for each document associated with each term layer; and (g) creating a z coordinate associated with each term layer for each x,y coordinate by applying a smoothing function to the x,y coordinates for each document, and superimposing upon one another all of the term layers." Contrary to the Examiner's assertion that Lantrip "clusters using centroids", Lantroid clusters (col. 2,

lines 39-43) and then determines the centroid coordinates of the clusters (col. 2, lines 43-45). These centroid coordinates are then used when creating a 3D display of the clusters (col. 2, lines 46-57). Nothing in Lantrip teaches or suggests the claimed feature of "clustering second documents in a second dataset using said centroid seeds, such that said second datset has a similar, based on said centroid seeds, clustering to that of said first dataset."

The Office Action acknowledges that "Lantrip fails to disclose generating a second vector space model by counting, for each word in said first dictionary, and number of said second documents in which said word occurs and classifying said second documents in said second vector space model using said first document classes to produce a classified second vector space model." Therefore, the Office Action provides that "col. 14, lines 28-36 of Ruocco indicate that vector clustering analysis may involve multiple datasets in order to gain the benefit of information analysis from multiple sources. It would have been obvious to one of ordinary skill in the art at the time of the invention to have vector clustering analysis involve multiple datasets in order to gain the benefit of information analysis from multiple sources." The Applicants respectfully disagree.

Ruocco, col. 14, lines 27-36 refers to a feature in claim 1of Ruocco, namely, "selecting a second electronic document and comparing the vector of the second electronic document with the first cluster vector to determine if the second document has similar characteristics, and assigning the second document vector to the first cluster vector if they have similar characteristics or designating the second document vector as a second cluster vector and assigning the second cluster vector to a second processor of the parallel processors if there are different characteristics." The Applicants submit that this feature does not equate to the claimed features of "wherein said second dataset comprises a new, but related ,based on said centroid seeds, dataset different than said first dataset". Rather this feature refers to that aspect of the Ruocco invention which uses cluster vectors (i.e., either the document vector if there is only one document in a cluster or the mathematical average of document vectors in a cluster of more than one document) when

determining whether or not to add a new document to an old cluster or to create a new cluster in a dataset.

Specifically, as discussed above, Ruocco discloses an improvement over legacy methods of clustering documents in a single dataset (e.g., a 32 document set, a 64 document set, a 128 document set (see col. 7, lines 30-35). Due to a determined need to organize a large number of documents in a document set, the Ruocco method forms clusters of similar documents, but instead of maintaining those clusters on a single processor, Ruocco maintains one or more clusters on multiple different processors. The initial cluster vector for a given cluster in Ruocco is disclosed as being the first document vector in the cluster. After that it is a mathematical average of document vectors in the cluster (see col. 14, line 25-26). As each new document is added to the dataset, the document is assigned to a particular cluster on a particular processor based on similarities between its document vector and the cluster vector of the particular cluster (see col. 14, lines 28-46). Thus, Ruocco uses cluster vectors (i.e., either the document vector if there is only one document in a cluster or the mathematical average of document vectors in a cluster of more than one document) when determining whether or not to add a new document to an old cluster or to create a new cluster in a dataset. However, nowhere in Ruocco does not it teach or suggest “)”wherein said second dataset comprises a new, but related ,based on said centroid seeds, dataset different than said first dataset.”

Therefore, the Applicants submit that the applied prior art references alone and in combination do teach or suggest the following features of independent claim 20: (1) “forming a first dictionary of most common words in a first dataset”; (2) “generating a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said word occurs”; (3) “clustering said first documents in said first dataset based on said first vector space model to produce first document classes”; (4) “generating a second vector space model by counting, for each word in said first dictionary, a number of said second documents in which said word occurs”; (5) “classifying said second documents in said second vector space model using said first document classes to produce a classified second vector space model”; (6) “determining a

mean of vectors in each class in said classified second vector space model to produce centroid seeds"; (7) "clustering second documents in a second dataset using said centroid seeds, such that said second dataset has a similar, based on said centroid seeds, clustering to that of said first dataset"; and (8)"wherein said second dataset comprises a new, but related ,based on said centroid seeds, dataset different than said first dataset." Furthermore, dependent claims 21-22 are similarly patentable, not only by virtue of their dependency from a patentable claim, but also by virtue of the additional features of the invention they define. In view of the foregoing, the Examiner is respectfully requested to reconsider and withdraw this rejection.

4. Independent Claim 23.

The Applicants submit that a prima facie case of obvious has not been established in support of rejecting claim 23 (See *In re Vaeck*, 947 F.2d 488, 20 USPQ2d 1438 (Fed. Cir. 1991)). Specifically, the cited prior art references as a whole do not suggest the desirability and, thus, the obviousness of making the combination (see *Hodosh v. Block Drug Co., Inc.*, 786 F.2d 1136, 1143 n.5, 229 USPQ 182, 187 n.5 (Fed. Cir. 1986)). Additionally, the combination of Ruocco and Lantrip does not teach or suggest all of the limitations of claim 23.

a. Lack Of Suggestion/Motivation To Combine.

The Applicants submit that the there is no suggestion or motivation, either in the references themselves or in the knowledge generally available to one of ordinary skill in the art, to combine the teachings of Lantrip and Ruocco.

Lantrip discloses a method of visually displaying the relative content of a large number of documents (see Abstract). Specifically, the relationships of the documents are presented in a 3D landscape with the relative sizes and heights of peaks in the landscape representing the relative significance of a relationship of an individual document in the document set to a topic or term (see col. 2, lines 26-31). As discussed in col. 2, lines 30-55, the Lantrip method includes the steps of creating a vector for each document in the set such that each vector represents the relative relationship of an

individual document to a term or topic. Then, the vectors are arranged into clusters. For each cluster, the "centroid coordinates" (i.e., the coordinates for the center of the cluster mass) are determined as well as the distance of each document in a cluster from the centroid. This information is ultimately used to generate the 3D display. Thus, Lantrip is only concerned with a single set of documents.

Ruocco discloses a processing system that utilizes parallel processors for organizing and clustering a large number of documents (see Abstract). Col. 4, lines 35-45, of Ruocco summarizes conventional document clustering in which text documents are converted to document vectors and clustered. Clusters of documents are represented by cluster vectors. As each new document is processed, its document vector is compared to the cluster vector for each cluster. If a document vector is similar to a given cluster vector, it is placed in that cluster. If it is not, a new cluster is formed. Col. 4, lines 45-60, summarizes the Ruocco method that uses multiple processors to store various clusters. In the Ruocco method, as each new document in a dataset is processed, its document vector is simultaneously compared to P cluster vectors, where P is the number of processors within the system. That is, the processors work simultaneously and each processor compares the document vector to each of its clusters.

The Office Action indicates that it "would have been obvious to one of ordinary skill in the art at the time of the invention to use the information contained in the centroid seeds from Lantrip for subsequent datasets as in Ruocco in order to improve analysis of subsequent datasets." However, as discussed above, Lantrip does not disclose "centroid seeds", but rather calculating centroid coordinates for clusters of documents in a single dataset in order to generate a 3D display of that dataset. Rocco on the other hand deals with a single dataset of documents and clustering those documents on different processors. Neither address how different, but related datasets, would be processed.

Specifically, the starting points for creating clusters in a dataset are typically selected in some random manner. In Ruocco, the starting point is based on the first document vector and the document vectors of any subsequently processed document that is not similar to already established clusters. In Landrip, the starting point is not

disclosed. In the present invention, however, it is recognized that clustering documents from new and different, but related datasets, results in a lack of continuity that is a drawback when one in interested in tracking changes in the data (trends) over time (see page 5, lines 0-9). This is a problem not addressed in either Ruocco or Landrip (as they disclose conventional techniques for selecting the starting points and no technique, respectively). The present invention solves this problem by intentionally biasing the clustering algorithm towards first classes in a first dataset by using centroid seeds generated from clusters corresponding to those first classes as the starting points for clustering a second new and different, but related, dataset. More specifically, the claimed invention solves the problem of finding new categories in a second data set that did not exist in the first data set, while at the same time maintaining as nearly as possible categories from the first data set as categories in the second data set. With the claimed invention, there is no requirement, and in fact it is not assumed, that the first and second datasets have any of the same data elements in them. They are allowed to have some of the same elements, but this is in no way a requirement for the claimed invention. The claimed invention is designed in such a way as to find the similarities between the two datasets, where they exist, while at the same time finding the key differences (emerging concepts) in the second dataset.

Nothing in either Ruocco or Lantrip suggests that subsequent datasets (related or not) would be processed any differently then stated in either reference, much less that they would be processed based on information from prior datasets (as in the present invention). Furthermore, nothing in Lantrip or Ruocco suggests that the centroid coordinates used to make the 3D display of Lantrip would be used for anything other than forming a 3D display, much less that they would be used as the seeds for forming the initial clusters for subsequent datasets to be processed according to the Ruocco invention (i.e., on multiple processors). Rather the only obvious Lantrip and Ruocco combination would result in the calculation of centroid coordinates for all of the clusters stored on all of the processors in of Ruocco in order to generate a 3D display as in Lantrip.

b. All Of The Claim Limitations Not Taught.

The Applicants further submit that neither Lantrip, nor Ruocco, teach or suggest the following features of the program storage device of independent claim 23: (1) "A program device readable by machine tangibly embodying a program of instructions executable by the machine to perform a method of clustering documents in datasets;" (2) "creating centroid seeds based on said first document classes"; (3) "clustering second documents in a second dataset using said centroid seeds, such that said second dataset has a similar, based on said centroid seeds, clustering to that of said first dataset"; (3) "wherein said second dataset comprises a new, but related ,based on said centroid seeds, dataset different than said first dataset"; and (5) "wherein said clustering of said first documents in said first dataset comprises: forming a first dictionary of most common words in said first dataset; generating a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said word occurs; and clustering said first documents in said first dataset based on said first vector space model."

In rejecting independent claim 23, the Office Action provided: "Regarding independent claim 23, , the applicant discloses the limitations substantially similar to those in claim 8. Claim 23 is similarly rejected." The Applicants respectfully disagree with the assertion that the claim limitations of claim 23 are substantially similar to that of claim 8. Claim 8 refers to a system and structural components thereof. Whereas, claim 23 refers to a program device readable by machine tangibly embodying a program of instructions executable by the machine to perform a method and sets out each of the individual method steps. Neither Lantrip, nor Ruocco, disclose such a program device. However, as discussed above, in rejecting claim 8, the Office Action referred, not to a system and the structural components thereof (as claimed), but rather to a method of clustering documents as well as several method steps several of which are set out in claim 23.

The Office Action provides that Lantrip discloses: "creating centroid seeds based on said first document classes (in col. 2, lines 43-45, the invention finds centroids)." The

Applicants respectfully disagree. As discussed above, the Applicants respectfully disagree. Col. 2, lines 43-45 of Lantrip refers to calculating "centroid coordinates" of clusters in a dataset (i.e., determining the coordinates for the center mass of each cluster after all the clusters are formed). These centroid coordinates of Lantrip are mapped points in a cluster that are used by Lantrip to ultimately generate a 3D display representing the dataset. They are not "centroid seeds" that are subsequently used as the starting points to create clusters for a new and different, but related, dataset.

The Office Action acknowledges that "Lantrip fails to disclose clustering second documents in a second dataset using said centroid seeds." Therefore, the Office Action cites col. 14, lines 10-45of Ruocco as disclosing "in the claim processing in parallel second datasets based on cluster information from previous cluster vectors (see col. 14, lines 28-30) in order to gain the benefit of information from previous clusters to improve analysis of subsequent datasets. Ruocco's invention further may be interpreted such that said second dataset has a similar clustering to that of said first dataset (as the term "similar" is sufficiently broad that any two given datasets would have some degree of similarity, see35 U.S.C. 112 rejection, above.), further wherein said second dataset comprises a new, but related dataset different than said first dataset (once the first dataset is transformed, it is by definition a new, but related dataset). It would have been obvious to one of ordinary skill in the art at the time of the invention to use the information contained in the centroid seeds from Lantrip for subsequent datasets as in Ruocco in order to improve analysis of subsequent datasets." The Applicant respectfully disagree.

Ruocco, col. 14, lines 10-45 refers to the all of claim 1of the invention of Ruocco Specifically, Ruocco, col. 14, lines 10-45 provides: " We claim: 1. In an arrangement of parallel processors in a computer information processing system, a parallel clustering method for examining preselected documents and grouping similar documents in the parallel processors for subsequent retrieval in an electronic digital format from the computer information processing system, the steps comprising: converting each preselected document into an electronic document in digital format; converting each electronic document into a vector, whereby a vector is a weighted list of the occurence of

different words and terms that appear in the document; selecting a first electronic document and designating the vector of the first electronic document as a first cluster vector whereby a cluster vector is the mathematical average of all of the document vectors having similar characteristics, and assigning the first cluster vector to a first processor of the parallel processors; selecting a second electronic document and comparing the vector of the second electronic document with the first cluster vector to determine if the second document vector has similar characteristics, and assigning the second document vector to the first cluster vector if they have similar characteristics or designating the second document vector as a second cluster vector and assigning the second cluster vector to a second processor of the parallel processors if there are different characteristics; and selecting each subsequent electronic document and comparing the vector of each subsequent electronic document with all existing cluster vectors simultaneously on each processor having a cluster vector, and assigning each subsequent document vector to a parallel processor having the most similar characteristics or designating the subsequent document vector as a subsequent cluster vector and assigning the subsequent cluster vector to a processor of the parallel processors if there are different characteristics.."

The Applicants submit that the neither the entire claimed invention in Ruocco, nor any portion thereof, equates to the features in claim 23 of "clustering second documents in a second dataset using said centroid seeds, such that said second dataset has a similar, based on said centroid seeds, clustering to that of said first dataset" and "wherein said second dataset comprises a new, but related, based on said centroid seeds, dataset different than said first dataset", as suggested by the Examiner. Rather the invention of Ruocco provides an improvement over legacy methods of clustering documents in a single dataset (e.g., a 32 document set, a 64 document set, a 128 document set (see col. 7, lines 30-35). Due to a determined need to organize a large number of documents in a document set, the Ruocco method forms clusters of similar documents, but instead of maintaining those clusters on a single processor, Ruocco maintains one or more clusters on multiple different processors. The initial cluster vector for a given cluster in Ruocco

is disclosed as being the first document vector in the cluster. After that it is a mathematical average of document vectors in the cluster (see col. 14, line 25-26). As each new document is added to the dataset, the document is assigned to a particular cluster on a particular processor based on similarities between its document vector and the cluster vector of the particular cluster (see col. 14, lines 28-46). Thus, Ruocco uses cluster vectors (i.e., either the document vector if there is only one document in a cluster or the mathematical average of document vectors in a cluster of more than one document) when determining whether or not to add a new document to an old cluster or to create a new cluster in a dataset. However, nowhere in Ruocco does not it teach or suggest "clustering second documents in a second dataset using said centroid seeds, such that said second dataset has a similar, based on said centroid seeds, clustering to that of said first dataset" and "wherein said second dataset comprises a new, but related, based on said centroid seeds, dataset different than said first dataset".

Finally, in rejecting claim 23, the Office Action fails to either indirectly by reference to claim 8 or directly address the feature of "wherein said clustering of said first documents in said first dataset comprises: forming a first dictionary of most common words in said first dataset; generating a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said word occurs; and clustering said first documents in said first dataset based on said first vector space model."

Therefore, the Applicants submit that the applied prior art references alone and in combination do teach or suggest the following features of independent claim 23: (1) "A program device readable by machine tangibly embodying a program of instructions executable by the machine to perform a method of clustering documents in datasets;" (2) "creating centroid seeds based on said first document classes"; (3) "clustering second documents in a second dataset using said centroid seeds, such that said second dataset has a similar, based on said centroid seeds, clustering to that of said first dataset"; (3) "wherein said second dataset comprises a new, but related ,based on said centroid seeds, dataset different than said first dataset"; and (5) "wherein said clustering of said first

documents in said first dataset comprises: forming a first dictionary of most common words in said first dataset; generating a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said word occurs; and clustering said first documents in said first dataset based on said first vector space model." Furthermore, dependent claims 26-29 are similarly patentable, not only by virtue of their dependency from a patentable claim, but also by virtue of the additional features of the invention they define. In view of the foregoing, the Examiner is respectfully requested to reconsider and withdraw this rejection.

B. CONCLUSION

In view the forgoing, the Board is respectfully requested to reconsider and withdraw the rejections of claims 8, 11-15, 17-23 and 26-29.

Please charge any deficiencies and credit any overpayments to Attorney's Deposit Account Number 09-0441.

Respectfully submitted,

Date: September 27, 2007

/Pamela M. Riley/
Pamela M. Riley, Esq.
Registration No. 40,146

Gibb & Rahman, LLC
2568-A Riva Road, Suite 304
Annapolis, MD, 21401
Voice: (410) 573-0227
Fax: (301) 261-8825
Customer No. 29154

IX. CLAIMS APPENDIX

1-7. (Cancelled).

8. A system for clustering documents in datasets comprising:

a storage having a first dataset and a second dataset;

a cluster generator operative to cluster first documents in said first dataset and produce first document classes;

a centroid seed generator operative to generate centroid seeds based on said first document classes;

a dictionary generator adapted to generate a first dictionary of most common words in said first dataset; and

a vector space model generator adapted to generate a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said word occurs,

wherein said cluster generator clusters said documents in said first dataset based on said first vector space mode,

wherein said cluster generator clusters second documents in said second dataset using said centroid seeds, such that said second dataset has a similar, based on said centroid seeds, clustering to that of said first dataset, and

wherein said second dataset comprises a new, but related, based on said centroid seeds, dataset different than said first dataset.

9-10. (Cancelled).

11. The system in claim 8, wherein said vector space model generator generates a second vector space model by counting, for each word in said first dictionary, a number of said second documents in which said word occurs.

12. The system in claim 11, further comprising a classifier adapted to classify said second documents in said second vector space model using said first document classes to produce a classified second vector space model and adapted to determine a mean of vectors in each class in said classified second vector space model, wherein said mean comprises said centroid seeds.

13. The system in claim 11, wherein:

said dictionary generator is adapted to generate a second dictionary of most common words in said second dataset,

said vector space model generator is adapted to generate a third vector space model by counting, for each word in said second dictionary, a number of said second documents in which said word occurs, and

said cluster generator is adapted to cluster said second documents in said second dataset based on said third vector space model to produce a second dataset cluster.

14. The system in claim 13, wherein said cluster generator is adapted to produce an adapted dataset cluster by clustering said second documents in said second dataset using said centroid seeds and said system further comprises:

a comparator adapted to compare classes in said adapted dataset cluster to classes in said second dataset cluster and add classes to said adapted dataset cluster based on said comparing.

15. A method of clustering documents in a first dataset having first documents and a related second dataset having second documents, said method comprising:

clustering said first documents to produce first document classes;

generating a vector space model of said second documents;

classifying said vector space model of said second documents using said first document classes to produce a classified vector space model; and

determining a mean of vectors in each class in said classified vector space model to produce centroid seeds; and

clustering said second documents using said centroid seeds, such that said second dataset has a similar, based on said centroid seeds, clustering to that of said first dataset,

wherein said second dataset comprises a new, but related, based on said centroid seeds, dataset different than said first dataset,

wherein said vector space model comprises a second vector space model and said clustering of said first documents in said first data comprises:

forming a first dictionary of most common words in said first dataset; and

generating a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said word occurs,

wherein said clustering of said first documents in said first dataset is based on said first vector space model.

16. (Cancelled).

17. The method in claim 16, wherein said generating of said second vector space model comprises counting, for each word in said first dictionary, a number of said second documents in which said word occurs.

18. The method in claim 17, further comprising:

forming a second dictionary of most common words in said second dataset;

generating a third vector space model by counting, for each word in said second dictionary, a number of said second documents in which said word occurs; and

clustering said documents in said second dataset based on said third vector space model to produce a second dataset cluster.

19. The method in claim 18, wherein said clustering of said second documents in said second dataset using said centroid seeds produces an adapted dataset cluster and said method further comprises:

comparing classes in said adapted dataset cluster to classes in said second dataset cluster; and

adding classes to said adapted dataset cluster based on said comparing.

20. A method of clustering documents in related datasets comprising:

forming a first dictionary of most common words in a first dataset;

generating a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said word occurs;

clustering said first documents in said first dataset based on said first vector space model to produce first document classes;

generating a second vector space model by counting, for each word in said first dictionary, a number of said second documents in which said word occurs;

classifying said second documents in said second vector space model using said first document classes to produce a classified second vector space model;

determining a mean of vectors in each class in said classified second vector space model to produce centroid seeds; and

clustering second documents in a second dataset using said centroid seeds, such that said second dataset has a similar, based on said centroid seeds, clustering to that of said first dataset,

wherein said second dataset comprises a new, but related ,based on said centroid seeds, dataset different than said first dataset.

21. The method in claim 20, further comprising:

forming a second dictionary of most common words in said second dataset;

generating a third vector space model by counting, for each word, in said second dictionary, a number of said second documents in which said word occurs; and

clustering said documents in said second dataset based on said third vector space model to produce a second dataset cluster.

22. The method in claim 21, wherein said clustering of said second documents in said second dataset using said centroid seeds produces an adapted dataset cluster and said method further comprises:

comparing classes in said adapted dataset cluster to classes in said second dataset cluster; and

adding classes to said adapted dataset cluster based on said comparing.

23. A program device readable by machine tangibly embodying a program of instructions executable by the machine to perform a method of clustering documents in datasets comprising:

clustering first documents in a first dataset to produce first document classes;

creating centroid seeds based on said first document classes; and

clustering second documents in a second dataset using said centroid seeds, such that said second dataset has a similar, based on said centroid seeds, clustering to that of said first dataset,

wherein said second dataset comprises a new, but related ,based on said centroid seeds, dataset different than said first dataset,

wherein said clustering of said first documents in said first dataset comprises:

forming a first dictionary of most common words in said first dataset;

generating a first vector space model by counting, for each word in said first dictionary, a number of said first documents in which said word occurs; and

clustering said first documents in said first dataset based on said first vector space model.

24-25. (Cancelled).

26. A program device readable by machine, tangibly embodying a program of instructions executable by the machine to perform said method in claim 25, said method further comprising generating a second vector space model by counting, for each word in said first dictionary, a number of said second documents in which said word occurs.

27. A program device readable by machine , tangibly embodying a program of instructions executable by the machine to perform said method in claim 26, wherein said creating of said centroid seeds comprises:

classifying said second vector space model using said first document classes to produce a classified second vector space model; and

determining a mean of vectors in each class in said classified second vector space model, wherein said mean comprises said centroid seeds.

28. A program device readable by machine, tangibly embodying a program of instructions executable by the machine to perform said method in claim 26, said method further comprising:

forming a second dictionary of most common words in said second dataset;

generating a third vector space model by counting, for each word in said second dictionary, a number of said second documents in which said word occurs; and

clustering said documents in said second dataset based on said third vector space model to produce a second dataset cluster.

29. A program device readable by machine, tangibly embodying a program of instructions executable by the machine to perform said method in claim 28, wherein said clustering of said second documents in said second dataset using said centroid seeds produces an adapted dataset cluster and said method further comprises:

comparing classes in said adapted dataset cluster to classes in said second dataset cluster; and

adding classes to said adapted dataset cluster based on said comparing.

30. (Cancelled).

X. EVIDENCE APPENDIX

There is no other evidence known to Appellants, Appellants' legal representative or Assignee which would directly affect or be directly affected by or have a bearing on the Board's decision in this appeal.

XI. RELATED PROCEEDINGS APPENDIX

There is no other related proceedings known to Appellants, Appellants' legal representative or Assignee which would directly affect or be directly affected by or have a bearing on the Board's decision in this appeal.